Nonparametric estimation of relative mortality from nested case-control studies

Ørnulf Borgan
Institute of Mathematics, University of Oslo,
P.O. Box 1053 Blindern,N-0316 Oslo 3, Norway

Bryan Langholz
Department of Preventive Medicine,
University of Southern California, School of Medicine,
2025 Zonal Ave, Los Angeles, California 90033-9987, U.S.A.

June 18, 1992

Abstract

Andersen *et al.* (1985, *Biometrics* **41**, 921-932) gave an estimator of the cumulative relative mortality comparing rates of death in a epidemiologic cohort to an external population as a function of time when covariate information is available on all cohort members. We present an analogous estimator when covariate information is known only on a nested case-control sample. Using counting process techniques, it is shown that this estimator is almost unbiased and an estimator of its variance is derived. Estimators of the relative mortality function, using kernel smoothing methods, and the average relative mortality over grouped time intervals are also presented. The methods are illustrated comparing rates of lung cancer mortality in a cohort of Montana smelter workers to that in the United States population.

1 Introduction

It has long been recognized that relative rates of mortality between subgroups of a large cohort may be easily and efficiently estimated using the nested case-control method of cohort sampling and standard conditional logistic regression analysis methods. One drawback to this method has been the lack of a reliable means of comparing cohort rates to those of an external population. While the pitfalls associated with the use of standardized mortality ratios (SMR) for this purpose have been well documented, they can prove useful in providing a sense of the difference between disease rates in the cohort and the general population, especially when there is little variation in exposure in the cohort and the external population is essentially unexposed. Andersen *et al.* (1985) gave partial likelihood methods for estimating relative mortality from full cohort data. Breslow and Langholz (1987, Appendix), discussed more fully in Breslow and Day (1987, Chapter 5), gave an estimator of the cumulative SMR as a function of time based on a nested case-control sample and found that, with a small number of controls, it is very biased. Using methods analogous to those developed by Borgan *et al.* (1992) for estimation in the proportional

[0] *Key words:* Case-control studies; Cohort studies; Epidemiology; Relative risk; Standardized mortality ratio

[0] *Abbreviated title:* Relative mortality estimation in case-control studies

hazards model, in this paper we describe estimation procedures for the relative mortality model of Andersen *et al.* (1985) and give an estimator of cumulative relative mortality which is almost unbiased. We illustrate the method's utility for SMR estimation by comparing lung cancer mortality in a cohort of Montana copper smelter workers to that in the United States population.

2 A model for nested case-control studies

Let the cohort under study consist of n individuals, indexed by $i = 1, 2, \ldots, n$, and denote by $\alpha_i(t) = \alpha_i(t, \mathbf{z}_i(t))$ the hazard rate at time t for the ith individual with covariates $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{ip}(t))^\mathsf{T}$. We will consider a Cox regression model for the relative hazard (Andersen *et al.*, 1985), which can be given as follows: Let $\mu_i(t)$ be the *known* hazard at time t for an individual from an external standard population corresponding to the ith individual (e.g. of the same sex and age as individual i), and let

$$\alpha_i(t) = \mu_i(t)\theta(t)\exp(\beta_0^\mathsf{T}\mathbf{z}_i(t)), \tag{1}$$

where $\theta(t)$ is an underlying *relative hazard* common to all individuals in the cohort, i.e. it is the relative rate between the baseline cohort subjects and the external population. If $\mathbf{z}$ is a measure of "exposure" and is rarely different from zero in the external population, then $\theta(t)$ is a measure of the extent to which $\mathbf{z}$ explains the excess risk in the cohort and may be interpreted as the SMR that would have been found had the cohort not been exposed.

Note that if we let $\mu_i(t) = 1$ for all i and t, (1) reduces to the classical proportional hazards model. Thus results for this model are obtained as special cases of those presented below.

Individuals are allowed to enter and leave the population under study, and we let $t_1 < t_2 < \cdots$ denote the times when failures are observed. We will assume that there are no tied failures, and let $\delta_{ij} = 1$ if the ith individual fails at t_j, $\delta_{ij} = 0$ otherwise. The risk set at time t is denoted $\mathcal{R}(t)$, and the number of individuals at risk at t is $n(t) = \#\mathcal{R}(t)$. We let $N_i(t) = \sum_{t_j \le t} \delta_{ij}$ be the process counting the number of observed failures for the ith individual in $[0, t]$, and let $Y_i(t)$ be an indicator process taking the value 1 if this individual is at risk "just before" time t and the value 0 otherwise. Thus $\mathcal{R}(t) = \{i : Y_i(t) = 1\}$ and $n(t) = \sum_{i=1}^n Y_i(t)$.

Under the usual assumption of independent censoring the *intensity process* $\lambda_i(t)$ of the counting process $N_i(t)$ is informally given by

$$\lambda_i(t)\mathrm{d}t = \mathrm{P}(\mathrm{d}N_i(t) = 1 \mid \mathcal{H}_{t-}) = \alpha_i(t)Y_i(t)\mathrm{d}t, \tag{2}$$

where $\mathrm{d}N_i(t) = N_i((t + \mathrm{d}t)-) - N_i(t-)$ is the increment of N_i over the small time interval $[t, t + \mathrm{d}t)$ and $\mathcal{H}_{t-}$ contains information about observed failures, entries, exits and changes in covariate values in the cohort up to, but not including, time t (e.g. Andersen and Borgan, 1985, Section 3; Andersen *et al.*, 1992, Section II.1).

In a nested case-control study (Thomas, 1977), one selects without replacement at each failure time t_j a random sample of controls of size $m - 1$ from the non-failing individuals at risk. We let $\widetilde{\mathcal{R}}(t)$ denote the sampled risk set at t were a failure to occur at that time. This will consist of the failing individual together with its sampled set of controls. As a technical point, the number of controls could also depend on time. Specifically, if $n(t) < m$ we would set $\widetilde{\mathcal{R}}(t) = \mathcal{R}(t)$ but, for simplicity of exposition, we will assume below that the

size of the sampled risk set does not depend on t. We let $\mathcal{F}_{t-}$ contain information about all observed events in the cohort as well as about the sampling of controls in $[0,t)$. Thus $\mathcal{F}_{t-}$ is $\mathcal{H}_{t-}$ augmented with the sampling information. Furthermore, we introduce $\mathcal{P}^{(m)}$ for the set of all subsets of $\{1,2,\ldots,n\}$ of size m. Then we have

$$\mathrm{P}\left(\widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \Delta N_i(t) = 1, \mathcal{F}_{t-}\right) = \binom{n(t)-1}{m-1}^{-1}, \tag{3}$$

for $\mathbf{r} \subset \mathcal{R}(t)$, $\mathbf{r} \in \mathcal{P}^{(m)}$, and $i \in \mathbf{r}$, where $\Delta N_i(t) = N_i(t) - N_i(t-)$ is the increment of N_i at t.

For $\mathbf{r} \in \mathcal{P}^{(m)}$ and $i \in \mathbf{r}$, we now define $N_{(i,\mathbf{r})}(t) = \sum_{t_j \le t} \delta_{ij} I(\widetilde{\mathcal{R}}(t_j) = \mathbf{r})$ as the number of times in $[0,t]$ the ith individual fails and the sampled risk set equals $\mathbf{r}$. Moreover, we assume that the nested case-control sampling is *independent* in the sense that the additional knowledge of which individuals have been sampled as controls before any time t do not alter the intensities of failures at t. Thus $\mathrm{P}(\mathrm{d}N_i(t) = 1 \,|\, \mathcal{F}_{t-}) = \mathrm{P}(\mathrm{d}N_i(t) = 1 \,|\, \mathcal{H}_{t-})$. Informally therefore, by (2) and (3), the intensity process $\lambda_{(i,\mathbf{r})}(t)$ of the counting process $N_{(i,\mathbf{r})}(t)$ is given by

$$\begin{aligned}
\lambda_{(i,\mathbf{r})}(t)\mathrm{d}t &= \mathrm{P}(\mathrm{d}N_{(i,\mathbf{r})}(t) = 1 \,|\, \mathcal{F}_{t-}) = \mathrm{P}(\mathrm{d}N_i(t) = 1, \widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \mathcal{F}_{t-}) \\
&= \mathrm{P}(\mathrm{d}N_i(t) = 1 \,|\, \mathcal{F}_{t-}) \mathrm{P}\left(\widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \Delta N_i(t) = 1, \mathcal{F}_{t-}\right) \\
&= \alpha_i(t) Y_i(t) \mathrm{d}t \binom{n(t)-1}{m-1}^{-1} I(\mathbf{r} \subset \mathcal{R}(t), \mathbf{r} \in \mathcal{P}^{(m)}, i \in \mathbf{r}).
\end{aligned}$$

These heuristics, combined with (1), imply that the counting processes $N_{(i,\mathbf{r})}(t)$ have intensity processes

$$\lambda_{(i,\mathbf{r})}(t) = Y_i(t)\mu_i(t)\theta(t)\exp(\beta_0^{\mathsf{T}}\mathbf{z}_i(t)) \binom{n(t)-1}{m-1}^{-1} I(\mathbf{r} \subset \mathcal{R}(t), \mathbf{r} \in \mathcal{P}^{(m)}, i \in \mathbf{r}). \tag{4}$$

By standard counting process theory (e.g. Andersen and Borgan, 1985, Section 3; Andersen *et al.*, 1992, Section II.4.1) it follows that

$$M_{(i,\mathbf{r})}(t) = N_{(i,\mathbf{r})}(t) - \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u, \tag{5}$$

for $\mathbf{r} \in \mathcal{P}^{(m)}$, $i \in \mathbf{r}$, are orthogonal (local) square integrable martingales. In particular the $M_{(i,\mathbf{r})}(t)$ are uncorrelated and have mean zero.

3 Estimation

Estimation of the regression parameter in (1) may for nested case-control studies be based on the partial likelihood

$$\mathcal{L}(\beta) = \prod_{t_j} \left\{ \frac{\mu_{i_j}(t_j)\exp(\beta^{\mathsf{T}}\mathbf{z}_{i_j}(t_j))}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} \mu_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{z}_l(t_j))} \right\}, \tag{6}$$

with i_j being the individual who fails at t_j (Oakes, 1981). The product in (6) is taken over all t_j and the same applies to the corresponding sums below. Note that we use β for the free parameter in (6), while the true value of the regression parameter is denoted β_0; cf. (1) and (4).

The vector of score functions may be written as

$$\begin{aligned}\mathbf{U}(\beta) &= \frac{\partial}{\partial\beta}\log\mathcal{L}(\beta) = \sum_{t_j}\sum_{i=1}^{n}\delta_{ij}\left\{\mathbf{z}_i(t_j) - \frac{\sum_{l\in\widetilde{\mathcal{R}}(t_j)}\mu_l(t_j)\mathbf{z}_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{z}_l(t_j))}{\sum_{l\in\widetilde{\mathcal{R}}(t_j)}\mu_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{z}_l(t_j))}\right\} \\ &= \int_0^\infty \sum_{\mathbf{r}\in\mathcal{P}^{(m)}}\sum_{i\in\mathbf{r}}\left\{\mathbf{z}_i(t) - \frac{\sum_{l\in\mathbf{r}}Y_l(t)\mu_l(t)\mathbf{z}_l(t)\exp(\beta^{\mathsf{T}}\mathbf{z}_l(t))}{\sum_{l\in\mathbf{r}}Y_l(t)\mu_l(t)\exp(\beta^{\mathsf{T}}\mathbf{z}_l(t))}\right\}\mathrm{d}N_{(i,\mathbf{r})}(t).\end{aligned}$$

From this the observed information matrix

$$\mathcal{I}(\beta) = -\frac{\partial^2}{\partial\beta^2}\log\mathcal{L}(\beta)$$

is easily derived.

Using (4) and (5) it is seen that the vector of score functions, evaluated at the true parameter value, equals the (vector valued) stochastic integral

$$\mathbf{U}(\beta_0) = \int_0^\infty \sum_{\mathbf{r}\in\mathcal{P}^{(m)}}\sum_{i\in\mathbf{r}}\left\{\mathbf{z}_i(t) - \frac{\sum_{l\in\mathbf{r}}Y_l(t)\mu_l(t)\mathbf{z}_l(t)\exp(\beta_0^{\mathsf{T}}\mathbf{z}_l(t))}{\sum_{l\in\mathbf{r}}Y_l(t)\mu_l(t)\exp(\beta_0^{\mathsf{T}}\mathbf{z}_l(t))}\right\}\mathrm{d}M_{(i,\mathbf{r})}(t). \quad (7)$$

Asymptotic properties of the estimator $\widehat{\beta}$, obtained by maximizing (6), can therefore be derived using results for counting processes, martingales and stochastic integrals in a similar manner as in Andersen and Gill (1982). In particular (under suitable regularity conditions), $n^{-1/2}\mathbf{U}(\beta_0)$ converges weakly, by the martingale central limit theorem, to a multivariate normal distribution with mean $\mathbf{0}$ and a certain covariance matrix Σ, while $n^{-1}\mathcal{I}(\beta^*)$ converges in probability to Σ for any β^* which is consistent for β_0. Therefore, by a standard argument using Taylor series expansions,

$$\sqrt{n}(\widehat{\beta}-\beta_0) = \Sigma^{-1}\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0) + o_p(1). \quad (8)$$

It follows that $\widehat{\beta}$ is asymptotically multivariate normally distributed around β_0 with a covariance matrix that may be estimated by $\mathcal{I}(\widehat{\beta})^{-1}$. Borgan *et al.* (1992) gave detailed proofs for nested case-control studies for the proportional hazards model along these lines which may be modified to accomodate model (1).

As an estimator for the integrated underlying relative hazard $\Theta(t) = \int_0^t\theta(t)\mathrm{d}t$ we suggest

$$\widehat{\Theta}(t;\widehat{\beta}) = \sum_{t_j\le t}\frac{1}{\sum_{l\in\widetilde{\mathcal{R}}(t_j)}(n(t_j)/m)\mu_l(t_j)\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{z}_l(t_j))}. \quad (9)$$

Note that in contrast to the raw estimator proposed by Breslow and Langholz (1987); cf. Section 4 below; no distinction is made in (9) between the case and its controls.

To motivate (9), we use (4) and (5) to write (9), with $\widehat{\beta}$ replaced by β_0, as

$$\widehat{\Theta}(t;\beta_0) = \sum_{t_j \le t} \sum_{i=1}^{n} \frac{\delta_{ij}}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j)\exp(\beta_0^{\mathsf{T}} \mathbf{z}_l(t_j))}$$

$$= \int_0^t \sum_{\mathbf{r} \in \mathcal{P}^{(m)}} \sum_{i \in \mathbf{r}} \frac{\mathrm{d}N_{(i,\mathbf{r})}(u)}{\sum_{l \in \mathbf{r}} (n(u)/m) Y_l(u)\mu_l(u)\exp(\beta_0^{\mathsf{T}} \mathbf{z}_l(u))}$$

$$= \int_0^t \sum_{\mathbf{r} \subset \mathcal{R}(u), \mathbf{r} \in \mathcal{P}^{(m)}} \frac{m}{n(u)} \binom{n(u)-1}{m-1}^{-1} \theta(u)\mathrm{d}u + \widehat{W}(t)$$

$$= \int_0^t \theta(u)\mathrm{d}u + \widehat{W}(t), \tag{10}$$

where

$$\widehat{W}(t) = \int_0^t \sum_{\mathbf{r} \in \mathcal{P}^{(m)}} \sum_{i \in \mathbf{r}} \frac{\mathrm{d}M_{(i,\mathbf{r})}(u)}{\sum_{l \in \mathbf{r}} (n(u)/m) Y_l(u)\mu_l(u)\exp(\beta_0^{\mathsf{T}} \mathbf{z}_l(u))} \tag{11}$$

is a (local) square integrable martingale. It follows that $\widehat{\Theta}(t;\beta_0)$ is unbiased for $\Theta(t)$, thereby giving a justification for the estimator (9). (We have here disregarded the possibility of empty risk sets; see the appendix for a discussion of this point.)

The $\theta(t)$ in (1) is the relative hazard between the baseline cohort subjects and the external population, and this is estimated by the slope of (9). It may also be of some interest to estimate the relative rate between an individual in the cohort with a specified covariate $\mathbf{z}_0$, fixed over time, and the reference population. This is $\exp(\beta_0^{\mathsf{T}}\mathbf{z}_0)$ times $\theta(t)$, and its integral $\Theta_{\mathbf{z}_0}(t) = \exp(\beta_0^{\mathsf{T}}\mathbf{z}_0)\Theta(t)$ is estimated by

$$\widehat{\Theta}_{\mathbf{z}_0}(t) = \exp(\widehat{\beta}^{\mathsf{T}}\mathbf{z}_0)\widehat{\Theta}(t;\widehat{\beta}). \tag{12}$$

As outlined in the appendix the asymptotic properties of (9) and (12) can be derived by arguments parallel to those in the proof of Theorem 3.4 in Andersen and Gill (1982) or, with more details spelled out, in Andersen *et al.* (1992, Theorem VII.2.3 and Corollaries VII.2.4-6). The result is that $\widehat{\Theta}_{\mathbf{z}_0}(t)$ is asymptotically normally distributed around $\Theta_{\mathbf{z}_0}(t)$. The covariance between $\widehat{\Theta}_{\mathbf{z}_0}(t)$ and $\widehat{\Theta}_{\mathbf{z}_0}(u)$ may be estimated by

$$\widehat{\sigma}^2(u,t) = \left\{\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{z}_0)\right\}^2 \tag{13}$$

$$\times \left\{\widehat{\omega}^2(u \wedge t) + \left(\widehat{B}(u;\widehat{\beta}) - \widehat{\Theta}(u;\widehat{\beta})\mathbf{z}_0\right)^{\mathsf{T}} \mathcal{I}(\widehat{\beta})^{-1} \left(\widehat{B}(t;\widehat{\beta}) - \widehat{\Theta}(t;\widehat{\beta})\mathbf{z}_0\right)\right\},$$

where

$$\widehat{\omega}^2(t) = \sum_{t_j \le t} \frac{1}{\left\{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j)\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{z}_l(t_j))\right\}^2}, \tag{14}$$

and

$$\widehat{B}(t;\beta) = \sum_{t_j \le t} \frac{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j) z_l(t_j) \exp(\widehat{\beta}^{\mathsf{T}} z_l(t_j))}{\left\{ \sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j) \exp(\widehat{\beta}^{\mathsf{T}} z_l(t_j)) \right\}^2}. \tag{15}$$

The results for the cumulative baseline relative hazard estimator (9) are obtained by inserting $z_0 = 0$ above.

From the estimator (9) for the integrated underlying relative hazard $\Theta(t)$, an estimator of the underlying relative hazard itself may be obtained by kernel function smoothing. To this end let the kernel function $K(x)$ be a bounded function that vanishes outside $[-1, 1]$, integrates to one and let bandwidth b be a positive parameter. The relative mortality estimator is then

$$\begin{aligned} \widehat{\theta}(t) &= b^{-1} \int_0^\infty K((t-s)/b) \mathrm{d}\widehat{\Theta}(s;\widehat{\beta}) \\ &= b^{-1} \sum_{t-b \le t_j \le t+b} K\left(\frac{t-t_j}{b}\right) \left\{ \sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j) \exp(\widehat{\beta}^{\mathsf{T}} z_l(t_j)) \right\}^{-1}. \end{aligned} \tag{16}$$

In practice, the bandwidth will be chosen by the investigator to control the smoothness and bias of the estimated curve. Under appropriate conditions, as in Andersen *et al.* (1992, Theorem VII.2.7), a suitable estimator of the variance of $\widehat{\theta}(t)$ is

$$\widehat{\tau}^2(t) = b^{-2} \sum_{t-b \le t_j \le t+b} K^2\left(\frac{t-t_j}{b}\right) \left\{ \sum_{l \in \widetilde{\mathcal{R}}(t_j)} (n(t_j)/m)\mu_l(t_j) \exp(\widehat{\beta}^{\mathsf{T}} z_l(t_j)) \right\}^{-2}; \tag{17}$$

see the appendix for further details. In our example, we will use the Epanechnikov kernel function $K(x) = 0.75(1 - x^2)$, $\mid x \mid \le 1$.

4 Example

We illustrate these methods by comparing lung cancer mortality rates in a cohort of 8014 Montana smelter workers who were exposed to various levels of arsenic compounds as part of the smelting process (Lee and Fraumeni, 1969, Lee-Feldstein, 1983) to those in the United States white male population tabulated as 5 year age and 5 year calendar period average rates (Breslow and Day, 1987, Appendix IIIc). These data and the substantive issues and results are discussed extensively in Breslow and Day (1987). In Section 5.5 of that text, expanding upon work of Breslow and Langholz (1987), they examined the SMR as a function of years since first employment, first in the full cohort and then with nested case-control samples of various sizes. Their "raw estimator" is similar in form to our $\widehat{\Theta}$ defined in (9) except that only the controls contribute to the denominator and, in our notation, m is replaced by $m - 1$. They found that this estimator is very biased and explored bias reduction methods to correct for this. We will estimate the cumulative relative mortality functions, without and with adjustment for covariates, based on these data using $\widehat{\Theta}$ given by (9). The factors we consider in the adjusted model are: date of hire

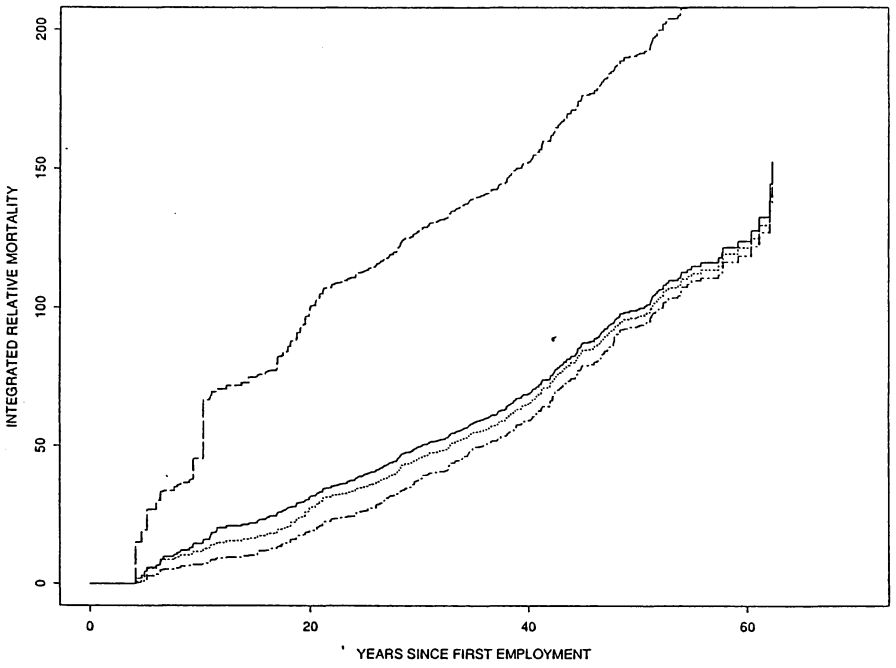

Figure 1. Estimates of the cumulative relative mortality function Θ for the unadjusted model: 1:1 matching (–·–·–·–·–·–) , 1:5 matching (- - - - - - - -), 1:100 matching (————), Breslow and Langholz raw estimator, 1:5 matching (– – – – – –).

(1885-1924 vs. 1925-1955) and years of employment in areas of the smelter considered to have heavy or moderate levels of arsenic exposure ($< 1, 1-4, 5-14, 15+$).

In order to reduce the computational burden, Breslow and Langholz (1987) rounded the years employed to integral years which resulted in 57 sampled risk sets with multiple failures in most sets. Rather than grouping the data as they did, we matched one and five controls to each of the 276 lung cancer cases so that there is only one case per matched set. This required arbitrarily breaking 8 truely tied failure times. Our 1:1 and 1:5 matched data sets had similar total sample sizes to Breslow and Langholz's 5 and 20 control matching. ($1 : m-1$ matching means that each sampled risk set of size m consists of 1 case and $m-1$ controls.) For comparison, we also generated a set with 100 controls per case which yields results similar to the full cohort.

Figure 1 shows estimated cumulative relative mortality functions for the unadjusted model. We also computed Breslow and Langholz's raw estimator from the 1:5 matched data set which is given for comparison. This estimator is clearly inferior to $\widehat{\Theta}$ based on either the 1:1 or 1:5 matched data sets. However, even these estimators appear to diverge from the 1:100 curve during the first 15 years after which they run reasonably parallel. This is very clearly apparent in Figure 2 which shows the kernel smoothed estimates (16) of the relative mortality using the Epanechnikov kernel function with a bandwidth of 7 years. The 1:1 and 1:5 matched data badly misrepresent the 1:100 data for the first 15 years since first employment, after which both run very close to the 1:100 data set curve. As we will explain in some detail below, this is not due to any inherent bias in the estimator, but may be attributed to the extreme skewness in the distribution of the external population rates for individuals at risk during this time interval. In any case, except for the first 15

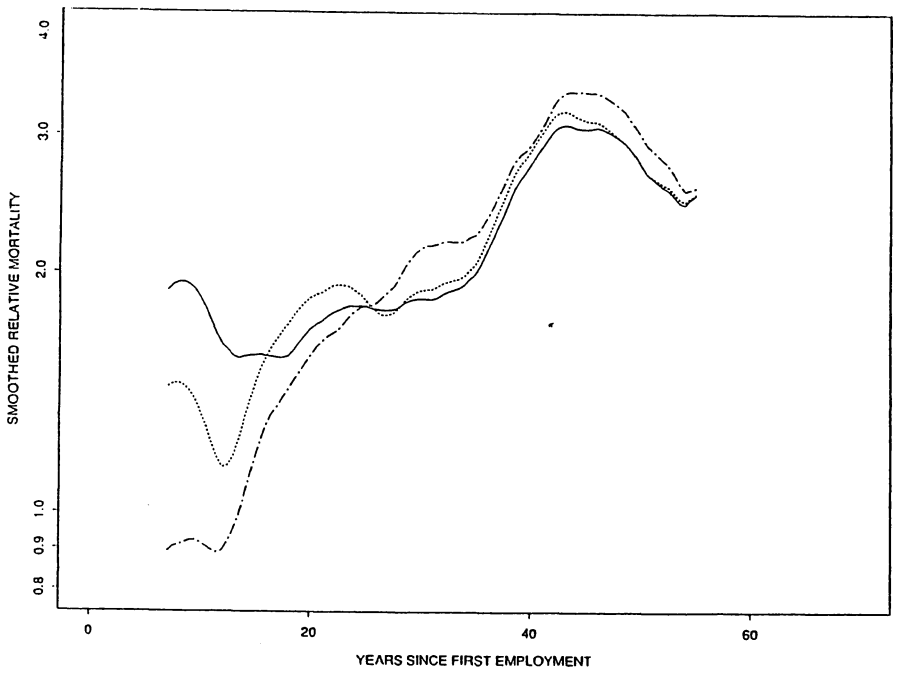

Figure 2. Kernel smoothed estimates of the relative mortality function θ for the unadjusted model: 1:1 matching (–·–·–·–·–·–), 1:5 matching (- - - - - - - -), 1:100 matching (————); bandwidth $b = 7$ years.

years, even the 1:1 matched data estimates the relative mortality quite well.

The estimates of the regression parameters for the adjusted model are given in Table 1. The estimates from the 1:5 matched data are reasonably close to those of the 1:100. Although the 1:1 matched data estimates are not quite so close, they retain the same qualitative pattern; the matching ratio is too low to expect high precision in the estimated parameters. Figure 3 shows unadjusted and adjusted $\widehat{\Theta}$ using the 1:5 matched data with 95% confidence intervals based on the log-transform (Bie *et al.*, 1987), i.e. $\widehat{\Theta}(t;\widehat{\beta})\exp(\pm 1.96\widehat{\sigma}(t,t)/\widehat{\Theta}(t;\widehat{\beta}))$, with $\widehat{\sigma}^2$ obtained from (13) with $\mathbf{z}_0 = 0$. The corresponding smoothed relative mortality functions with log-transformed confidence intervals based on (16) and (17) are shown in Figure 4. Rather than show similar curves for the

Table 1: *Estimated regression coefficients and standard errors for various matching ratios.*

	Matching ratio			
Covariate	1:1	1:5	1:100	1:1 Stratified[a]
Pre-1925 employment	0.525 (.265)	0.659 (.198)	0.763 (.174)	0.528 (.265)
Duration of heavy/ moderate arsenic exposure				
0-1 yr	1 (-)	1 (-)	1 (-)	1 (-)
1-4 yr	0.524 (.264)	0.706 (.200)	0.673 (.164)	0.560 (.262)
5-14 yr	0.208 (.334)	0.532 (.259)	0.479 (.214)	0.181 (.330)
15+ yr	1.194 (.332)	1.108 (.224)	0.923 (.181)	1.198 (.332)

[a]Only the 1-14 year risk sets were stratified, the others are from the 1:1 matched data.

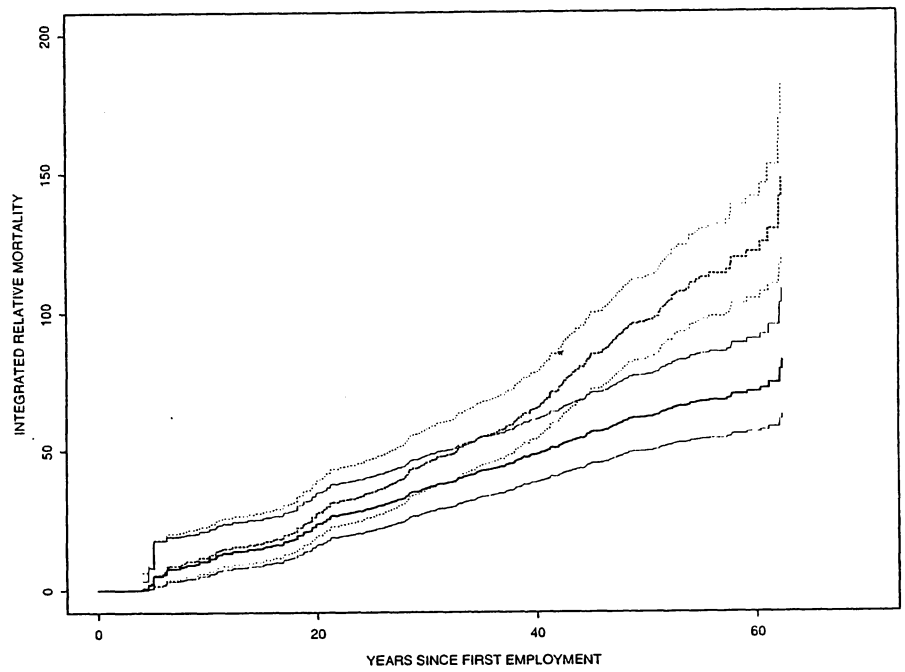

Figure 3. Estimates of the cumulative relative mortality function Θ for unadjusted (- - - - - - - -) and adjusted (————) models with 95% confidence intervals; 1:5 matching.

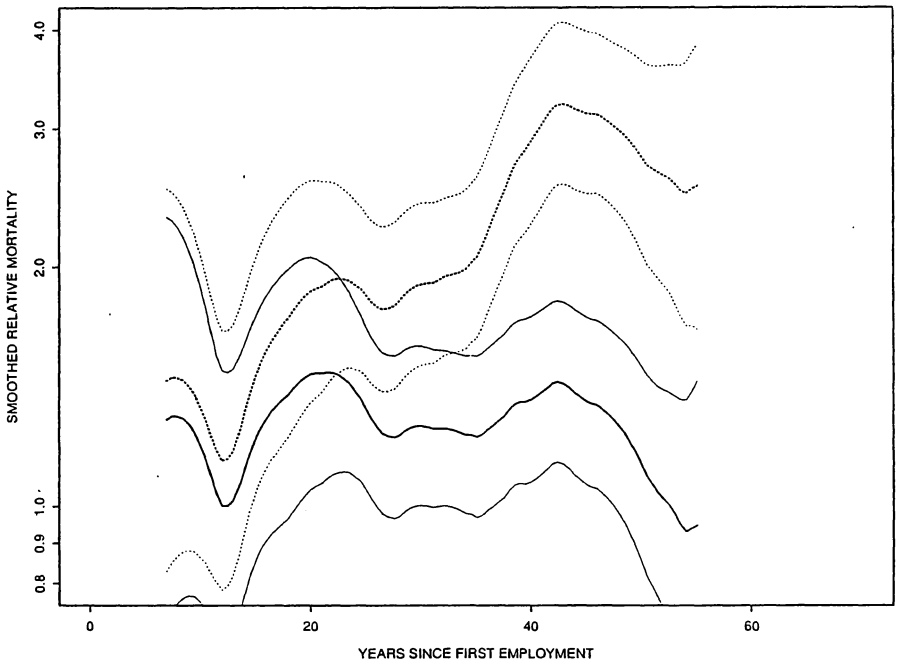

Figure 4. Kernel smoothed estimates of the relative mortality function θ for the unadjusted (- - - - - - - -) and adjusted (————) models with 95% confidence intervals; 1:5 matching, bandwidth $b = 7$ years.

Table 2: *Unadjusted and adjusted SMRs (standard errors) over grouped time intervals.*

Years since first employment	Matching ratio			
	1:1	1:5	1:100	1:1 Stratified[a]
		Unadjusted model		
1-14 yr	0.68 (.179)	1.10 (.285)	1.46 (.315)	1.55 (.369)
15-29 yr	1.86 (.247)	1.94 (.238)	1.86 (.192)	1.86 (.247)
30-59 yr	2.68 (.270)	2.53 (.246)	2.47 (.237)	2.68 (.270)
		Adjusted model		
1-14 yr	0.63 (.175)	0.98 (.269)	1.28 (.280)	1.36 (.333)
15-29 yr	1.46 (.230)	1.46 (.204)	1.42 (.166)	1.45 (.229)
30-59 yr	1.34 (.301)	1.16 (.201)	1.09 (.176)	1.34 (.300)

[a]Only the 1-14 year risk sets were stratified, the others are from the 1:1 matched data.

other matching ratios, in Table 2,

we give SMRs over grouped time intervals (1-14, 15-29, 30-59 years) as is typically done for full cohort data. These are computed as the average slopes of the $\widehat{\Theta}$ over the actual intervals (and therefore deviate slightly from what is ordinarily understood by SMRs). Thus we have calculated the SMR over an interval $[t_1, t_2]$ as

$$\mathrm{SMR}(t_1, t_2) = \frac{\widehat{\Theta}(t_2;\widehat{\beta}) - \widehat{\Theta}(t_1;\widehat{\beta})}{t_2 - t_1}, \tag{18}$$

and its standard error may by (13) (with $\mathbf{z}_0 = 0$) be estimated as

$$(t_2-t_1)^{-1}\left\{\widehat{\omega}^2(t_2) - \widehat{\omega}^2(t_1) + \left(\widehat{B}(t_2;\widehat{\beta}) - \widehat{B}(t_1;\widehat{\beta})\right)^{\mathsf{T}} \mathcal{I}(\widehat{\beta})^{-1}\left(\widehat{B}(t_2;\widehat{\beta}) - \widehat{B}(t_1;\widehat{\beta})\right)\right\}^{1/2}. \tag{19}$$

From the 1:100 matched data results in Table 2, it is evident that the covariates, as we have modelled them, explain much of the excess lung cancer mortality in the cohort though there is still a 10-40% excess which is not explained. We speculate that this is due to higher smoking levels in the smelter workers than in the general population. The SMRs computed from the 1:1 and 1:5 matched data sets are quite close to the 1:100 data for the 15-29 and 30-59 years since first employment categories. Striking, however, is how poorly the 1-14 years category is represented by the sampled data. The 1:1 and 1:5 matched data badly misrepresent the 1:100 matched data SMR for this category.

Although one might prefer an example which simply illustrates that our methods perform as expected, it is instructive to explore the reason for the extreme behavior of the sampled data estimator observed in the 1-14 year period. For the sake of this discussion we will consider the unadjusted model so that the sums in the denominators of (9) consist only of the external population rates determined by the age and calendar-year associated with each subject in the sampled risk set. The external rates vary greatly, especially across age, with 1,000 fold differences in rates between 20 year olds and 60 year olds. Because the time scale used in the analysis is duration since first employment, men of varying ages contribute to each risk set. There is the constraint that a subject must be somewhat older than the number of years employed so that ages, and thus the external rates for individuals at risk, become more homogeneous with time. However, during the first 10 to 15 years of employment, the vast majority of men are young and have very low rates, but the cases typically come from the 40-50 year olds who are at much higher risk. Figure 5 shows the distribution of the population rates for the risk sets over time. The extreme

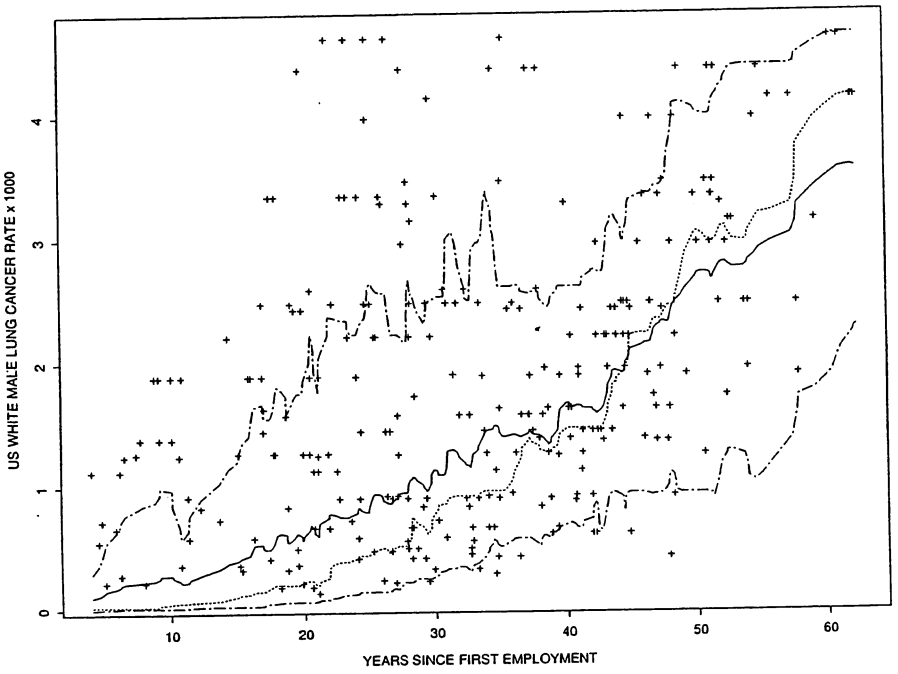

Figure 5. Smoothed estimates of the mean (———), median (- - - - - - - -), 10th and 90th percentiles (–·–·–·–·–·–) of the distributions of population rates in subjects at risk by time since first employment. Case rate values are indicated by plus signs (+). These were estimated from the 1:100 matched set.

skewness during the 1-14 year period is readily apparent. The plusses in Figure 5 are the population rates associated with the cases. Of the 24 cases which occurred during the first 15 years, all are well above the median rate.

At this point, it is necessary to develop a heuristic understanding of why the estimator (9) is (almost) unbiased. Consider 1:1 matching and, for simplicity, suppose that the controls all have the rate value of approximately 0.00003 per person year, approximately the median rate during the first 15 years of employment, and that the risk sets have around 5000 subjects. Cases with large rates, say 0.001, result in a small jump in $\widehat{\Theta}$ of about 0.4. A case with the median rate value would result in a jump size of about 7; a case with the first quartile value would double this. Since there are proportionally more cases with high rates, the jump sizes of $\widehat{\Theta}$ "average out" because the many small jumps are compensated for by an occasional large jump when a low risk case occur. Based on the population rates in the present study, about one low risk case (with rate less than the median) was expected during the first 15 years, and such a case could have corrected the deficit observed in the sampled data set SMRs of Table 2. Thus, the crux of the problem is that of small sample size akin to the problems of estimating a small probability with a small sample. It is a problem of variability and skewness of the estimator and not one of bias. Unfortunately, completely analogous to the small probability situation, the estimated variance fails to capture the true variability in the estimator. Interestingly, the standard error estimates for the 1-14 years SMRs in Table 2 actually increase with matching ratio.

The effect of a higher matching ratio is to reduce the difference in the jump size of $\widehat{\Theta}$ between low and high risk cases; in the full cohort the jump size is independent of

the case's rate. However, in this extreme situation, the matching ratio needs to be very large indeed to dampen the effect of the case. On the other hand, it does provide us with the opportunity to illustrate how stratified nested case-control sampling (Langholz and Borgan, 1992), a recently developed generalization of the simple nested case control design, provides a solution to this problem. In this sampling design, m_l subjects are picked from the $n_l(t_j)$ subjects in sampling stratum l at failure time t_j (and the failure is assumed to be sampled with probability one). The results of Section 3 continue to hold for this design provided that the weights $n(t_j)/m$ in (9) and (14) – (17) are replaced by $n_l(t_j)/m_l$ for subjects sampled from stratum l and the same weights are inserted in the partial likelihood (6). We illustrate this approach with 1:1 matched stratified sampling. Since the problem only exists during the first 15 years, stratified sampling was only carried out for the 24 1-14 years risk sets with the control randomly sampled from subjects with rates in the lower 90th percentile of the distribution if the case value was from the upper 10th percentile and reverse if the case value was from the lower 90th percentile. The weights used in the partial likelihood and in $\widehat{\Theta}$ for these risk sets were then $.9 \times n(t_j)$ and $.1 \times n(t_j)$ for the subject from the lower 90th percentile and upper 10th percentile, respectively. The other sampled risk sets were from the 1:1 matched (simple) nested case-control sample used above. The results are given in the last columns of Tables 1 and 2. It should not be surprising that there is little change from the 1:1 matched estimates in the second column of Table 1. Since the study began in 1938, very few of the subjects with less than 15 years since first employment could have been employed before 1925 nor could they have acquired 15+ years of arsenic exposure. Further, because there is a delay of some years between exposure and increased risk due to that exposure, i.e., a latency effect, it is unlikely that many of these cases were the result of arsenic exposure in the smelter. However, as seen in the last column of Table 2, stratification greatly improves the estimates of the SMR for the 1-14 years interval; they are much closer to those of 1:100 matched set.

5 Discussion

With the methodology we have provided, a nested case-control sample may be used to estimate the relative mortality function comparing the mortality rates in an external population to rates for those in a cohort with a covariate value $\mathbf{z}_0$. The estimators and their associated variances are easy to compute, they are simply functions of the sum of weighted relative risks from the sampled risk sets. Algorithms for estimating these quantities for the full cohort may be easily modified to accomodate nested case-control sampled data, the only difference is that the contribution from a sampled risk set is weighted by $n(t_j)/m$. Such curves may be summarized by grouped time SMRs as in Table 2 with estimates and standard errors given (for baseline subjects) by (18) and (19). Though we have refered to single failure times, these results hold without modification for multiple event data. Also, many other hazard structures of the form

$$\alpha_i(t) = \theta(t) f(\mu_i(t), \beta_0^{\mathsf{T}} \mathbf{z}_i(t))$$

may be accomodated in an obvious way.

As in the classical Cox proportional hazard model, estimation of the baseline hazard based on data with tied failure times poses real difficulties. Fortunately, the time of diagnosis or death of cohort members is generally quite accurately known so that there

are few ties. We suggest that these be randomly broken and that the risk sets be formed as if this were the true ordering. Sampling, and subsequent analysis of the data, would then proceed as if there were no ties.

The methods we have presented extend in an obvious way to general nested case-control sampling schemes (Borgan *et al.*, 1992) requiring only that appropriate weights are used in (9) and (14) – (17) as well as in the partial likelihood (6). This was illustrated for stratified nested case control sampling in the Montana smelter workers example.

The Montana smelter workers data may be an unfortunate choice of an example in that it may leave some readers sceptical of the validity of our methods. We point out that, except for the 1-14 year interval, the methods performed very well; even the 1:1 matched data set is adequate. This is representative of behavior one may typically expect. The poor behavior in the 1-14 year period is due to the extreme skewness of the distribution of population rate values. This phenomenon would also be expected in estimation of the baseline hazard in the classical proportional hazards model (i.e. $\mu_i(t) = 1$ for all i and t in (1)) when the variation of relative risks within the cohort is large and the covariate distributions highly skewed. There is well know parallel behavior of regression parameter estimates in that situation; efficiency as a function of relative risk drops of precipitously when "exposure" is rare (Breslow *et al.*, 1983). In any of these situations, stratified nested case-control sampling may provide an efficient alternative when there is some (possibly crude) measure of risk available on most of the cohort members on which stratification can by based.

Acknowledgements

This research was done while both authors were on sabbatical leave at the MRC Biostatistics Unit, Cambridge, England, the academic year 1991/92. The MRC Biostatistics Unit is acknowledged for its hospitality and for providing us with the best working facilities during this year. Ørnulf Borgan has been supported by the Norwegian Research Council for Science and the Humanities. Bryan Langholz has been supported by National Cancer Institute grant CA14089. The authors also thank Edward Rappaport for programming assistance.

References

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N., and Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics*, **41**, 921–932.

Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: A review (with discussion). *Scandinavian Journal of Statistics*, **12**, 97–158.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1992). *Statistical models based on counting processes.* Springer Verlag, New York. (in press).

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statists*, **10**, 1100–1120.

Bie, O., Borgan, Ø., and Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, **14**, 221–233.

Borgan, Ø., Goldstein, L., and Langholz, B. (1992). Generalized nested case-control sampling in Cox's regression model: a marked point process approach. *in preparation.*

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research.* Volume 2 – *The Design and Analysis of Cohort Studies, IARC Scientific Publications,* Vol. 82. International Agency for Research on Cancer, Lyon.

Breslow, N. E. and Langholz, B. (1987). Nonparametric estimation of relative mortality functions. *Journal of Chronic Diseases,* **131**, (Suppl. 2), 89S–99S.

Breslow, N. E., Lubin, J. H., Marek, P., and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association,* **78**, 1–12.

Langholz, B. and Borgan, Ø. (1992). Stratified nested case-control sampling in the Cox regression model. *in preparation.*

Lee, A. and Fraumeni, J. (1969). Arsenic and respiratory cancer in man: an occupational study. *Journal of the National Cancer Institute,* **42**, 1045–1052.

Lee-Feldstein, A. (1983). Arsenic and respiratory cancer in humans: follow-up of copper smelter employees in montana. *Journal of the National Cancer Institute,* **70**, 601–610.

Oakes, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *International Statistical Review,* **49**, 235–264.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society A,* **140**, 469–491.

Appendix

The large sample properties of the estimators (9) and (12) for the integrated relative hazard can be derived by arguments parallel to those in the proof of Theorem 3.4 in Andersen and Gill (1982), and as spelled out in more detail by Andersen *et al.* (1992, Theorem VII.2.3 and Corollaries VII.2.4-6).

We first mention a technical point which was suppressed in Section 3. In (10) and (11) we have disregarded the possibility that the denominators may become zero, i.e. the possibility of empty risk sets. A careful treatment of this problem requires the introduction of an indicator variable (which is zero when the risk set is empty) in the numerators in these formulas, cf. the above mentioned references. However, for large sample purposes we may safely ignore this problem.

We now use a Taylor series expansion to get

$$\widehat{\Theta}(t;\widehat{\beta}) = \widehat{\Theta}(t;\beta_0) - (\widehat{\beta}-\beta_0)^{\mathsf{T}}\widehat{B}(t;\beta^*), \tag{A.1}$$

where $\widehat{B}(t;\beta)$ is given by (15), and β^* is on the line segment joining $\widehat{\beta}$ and β_0. It can be shown as in the works mentioned above that $\widehat{B}(\cdot\,;\beta^*)$ converges uniformly in probability to a certain deterministic function $B(\cdot\,;\beta_0)$. It follows by (10) and (A.1) that the processes

$$\sqrt{n}\left(\widehat{\Theta}(\cdot\,;\widehat{\beta}) - \Theta(\cdot)\right) - \sqrt{n}(\widehat{\beta}-\beta_0)^{\mathsf{T}}B(\cdot\,;\beta_0) \tag{A.2}$$

and $\sqrt{n}\,\widehat{W}(\cdot)$ asymptotically have the same distribution. Thus, by the martingale central limit theorem, the asymptotic distribution of (A.2) is that of a mean zero Gaussian martingale. The variance function of this limiting process may be estimated uniformly consistently by $n\widehat{\omega}^2(\cdot)$ (cf. (14)). Furthermore, the predictable covariation process between the martingale $\mathbf{U}_t(\beta_0)$, obtained by replacing ∞ by t in (7), and the martingale (11) is zero, so that the processes (A.2) and $n^{-1/2}\mathbf{U}.(\beta_0)$ are asymptotically independent. Therefore, by (8), the process (A.2) is asymptotically independent of $\sqrt{n}(\widehat{\beta}-\beta_0)$. As a consequence of this the process $\sqrt{n}(\widehat{\Theta}(\cdot\,;\widehat{\beta}) - \Theta(\cdot))$ converges weakly to a mean zero Gaussian process with a covariance function which may be estimated uniformly consistently by

$$n\widehat{\omega}^2(u\wedge t) + n\widehat{B}(u;\widehat{\beta})^{\mathsf{T}}\mathcal{I}(\widehat{\beta})^{-1}\widehat{B}(t;\widehat{\beta}).$$

It also follows that the asymptotic covariance of $\sqrt{n}(\widehat{\beta} - \beta_0)$ and $\sqrt{n}(\widehat{\Theta}(t;\widehat{\beta}) - \Theta(t))$ may be estimated uniformly (in t) consistently by

$$-n\mathcal{I}(\widehat{\beta})^{-1}\widehat{B}(t;\widehat{\beta}).$$

Finally, by a Taylor series expansion, the processes

$$\sqrt{n}(\widehat{\Theta}_{\mathbf{z}_0}(\cdot) - \Theta_{\mathbf{z}_0}(\cdot)) \tag{A.3}$$

and

$$\exp(\beta_0^{\mathsf{T}}\mathbf{z}_0)\sqrt{n}\left(\widehat{\Theta}(\cdot\,;\widehat{\beta}) - \Theta(\cdot)\right) + \exp(\beta_0^{\mathsf{T}}\mathbf{z}_0)\Theta(\cdot)\mathbf{z}_0^{\mathsf{T}}\sqrt{n}(\widehat{\beta} - \beta_0)$$

asymptotically have the same distribution. Therefore (A.3) converges weakly to a mean zero Gaussian process with a covariance function which may be estimated uniformly consistently by $n\widehat{\sigma}^2(u,t)$ (cf. (13)). In particular for a fixed value of t, $\widehat{\Theta}_{\mathbf{z}_0}(t)$ is asymptotically normally distributed as described in Section 3.

We will study the asymptotic distribution for the kernel function estimator (16) for kernel functions which satisfy

$$\int_{-1}^{1} K(t)\mathrm{d}t = 1, \quad \int_{-1}^{1} tK(t)\mathrm{d}t = 0 \text{ and } \int_{-1}^{1} t^2K(t)\mathrm{d}t = k_2 > 0, \tag{A.4}$$

which is the case for all symmetric nonnegative functions with unit integral (like the Epanechnikov kernel function).

In order to derive asymptotic results one has to assume that the bandwidth $b = b_n$ tends to zero at a certain rate as $n \to \infty$. Assuming the bandwidth to be of the order $n^{-1/5}$, and moreover that $\alpha_0(u)$ is twice continuously differentiable in a neigbourhood of t, we may use (10), (A.4) and (A.1) to get

$$\begin{aligned}
&\sqrt{nb}\left(\widehat{\theta}(t) - \theta(t) - \frac{1}{2}b^2\theta''(t)k_2\right) \\
&= \sqrt{nb}\left\{b^{-1}\int_0^\infty K((t-u)/b)\,(\theta(u) - \theta(t))\,\mathrm{d}u - \frac{1}{2}b^2\theta''(t)k_2\right\} \\
&\quad + \sqrt{(n/b)}\int_0^\infty K((t-u)/b)\mathrm{d}\widehat{W}(u) \\
&\quad - \sqrt{nb}(\widehat{\beta} - \beta_0)^{\mathsf{T}}b^{-1}\int_0^\infty K((t-u)/b)\mathrm{d}\widehat{B}(u;\beta^*)
\end{aligned} \tag{A.5}$$

as in the proof of Theorem VII.2.7 in Andersen *et al.* (1992). Here the first and third term on the right hand side of (A.5) converges in probability to zero, while the second is a stochastic integral with respect to the martingale (11). It follows, using the martingale central limit theorem, that

$$\sqrt{nb}\left(\widehat{\theta}_0(t) - \theta(t) - \frac{1}{2}b^2\theta''(t)k_2\right)$$

converges weakly to a normal distribution with mean zero and a variance that can be consistently estimated by $nb\widehat{\tau}^2(t)$ (cf. (17)). Thus $\widehat{\alpha}_0(t)$ is asymptotically normally distributed with an asymptotic bias $\frac{1}{2}b^2\alpha_0''(t)k_2$ and with a variance that can be estimated by (17) as stated in Section 3.